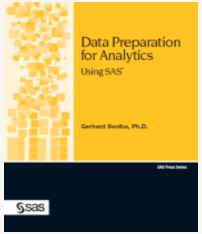
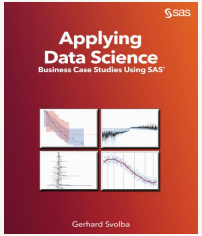
Applying Data Science – Ten Things Advanced Analytics and Data Science can do for your business

Gerhard Svolba, Analytic Solution Architect
SAS Austria

Amsterdam, October 17th, 2017





ANALYTICS EXPERIENCE 2017

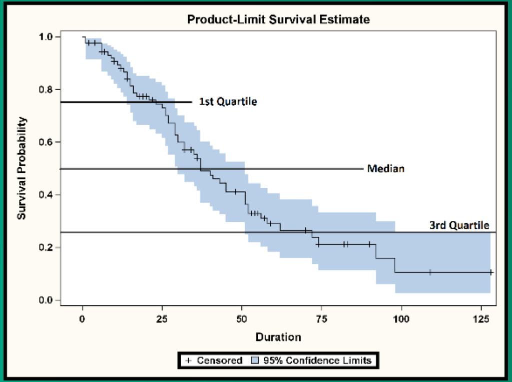
Agenda

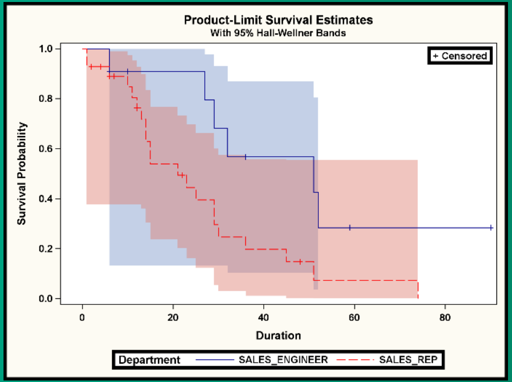
- 10 times „Data Science in Action“
  - Supervised Machine Learning Methods
  - Unsupervised Machine Learning Methods
  - Simulations
- Data Science and Advanced Analytics with the SAS Analytic Platform
- Summary and Links

Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention

Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?

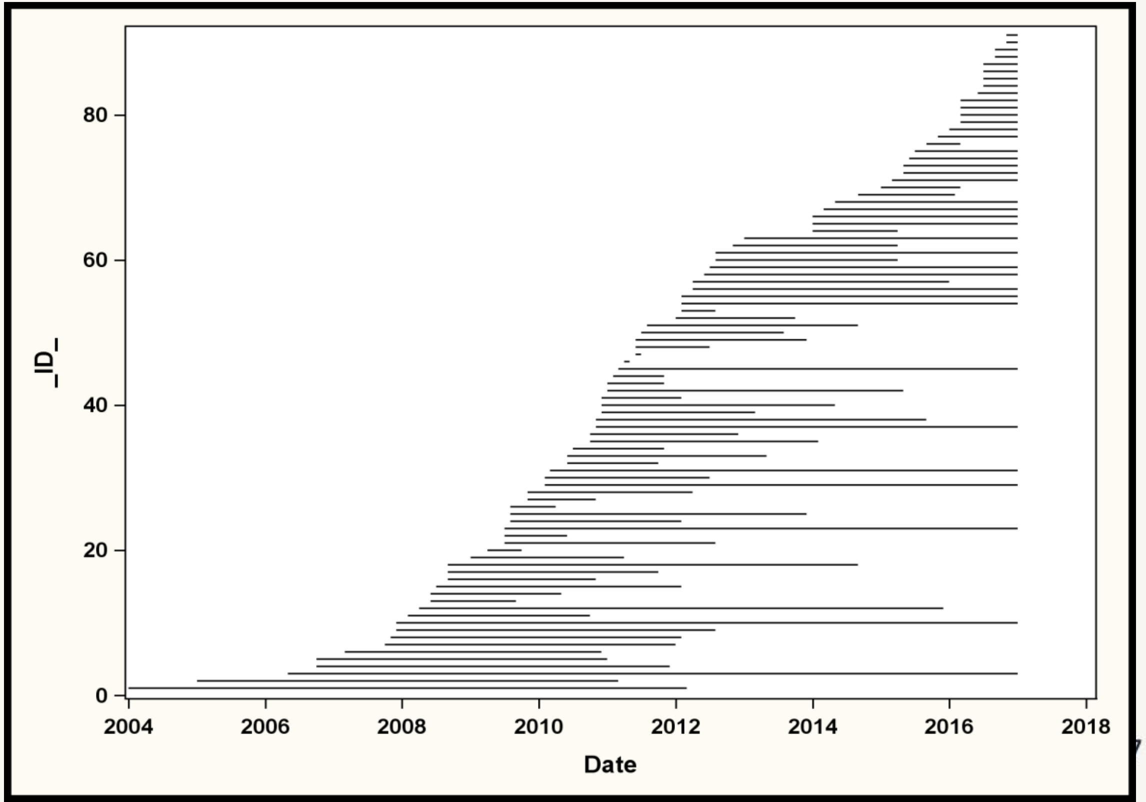




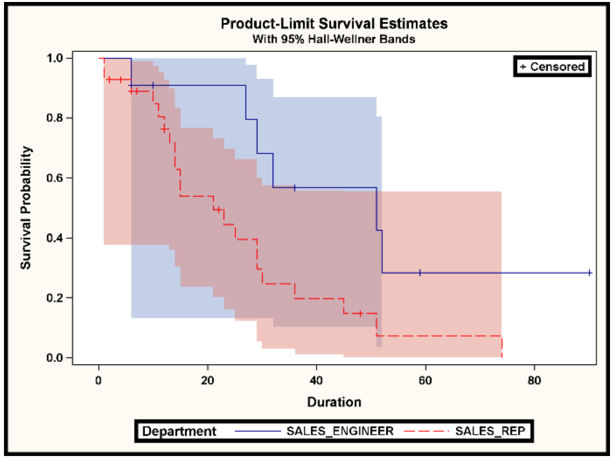
Survival analysis methods: Kaplan-Meier estimates Cox Proportional Hazards regression
Survival Data Mining

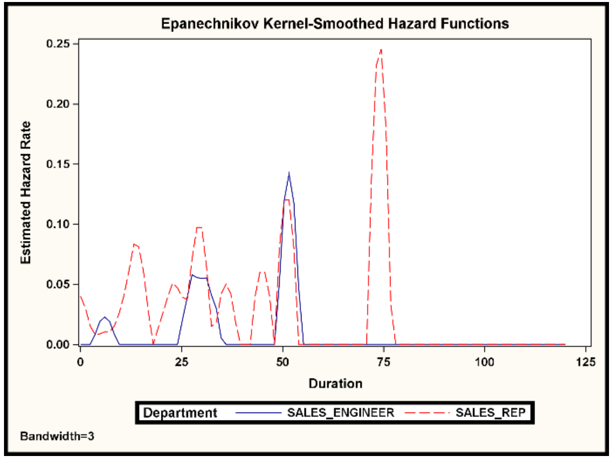
We do not have an event date for all employees (luckily ☺)!

- Observe Careers per Employee
  - Different length
  - „Left company“ or „censored“



The Kaplan Meier Method and the Cox Proportion Hazards Regression can deal with Missing Endpoints





Kaplan Meier Methods and Cox Proportional Hazards Regression:
Sales engineers have a better survival time than sales representatives.

Studying the Hazard Curves:
There is high risk to lose your sales engineers after 26 and after 50 months.

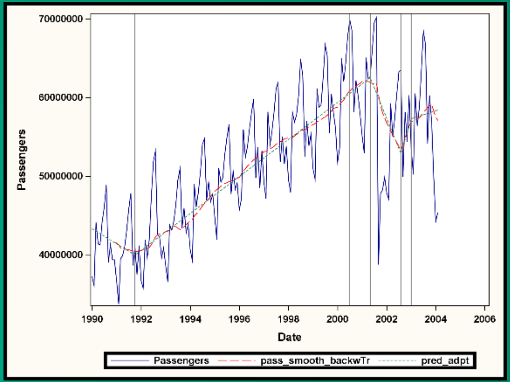
How long will Gerhard still stay in our company?

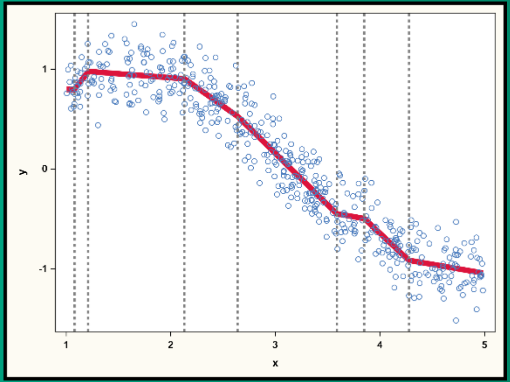
Given certain risk factors, what is the expected survival in 6 months and the probability to resign within the next 6 months.

EmpNo	Department	Gender	TechKnowH...	_T_	EM_SURVFCST	EM_SURVEVENT	T_FCST
1003	TECH_SUPPORT	M	YES	128	0.240	0.000	134
1010	TECH_SUPPORT	M	YES	109	0.240	0.011	115
1023	SALES_ENGINEER	M	YES	90	0.108	0.313	96
1029	TECH_SUPPORT	M	YES	83	0.386	0.133	89
1031	TECH_SUPPORT	F	YES	82	0.177	0.219	88
1037	ADMINSTRATION	M	NO	74	0.471	0.066	80
1045	ADMINSTRATION	M	NO	70	0.494	0.053	76
1054	TECH_SUPPORT	F	YES	59	0.316	0.102	65
1055	SALES_ENGINEER	M	YES	59	0.313	0.103	65

Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data



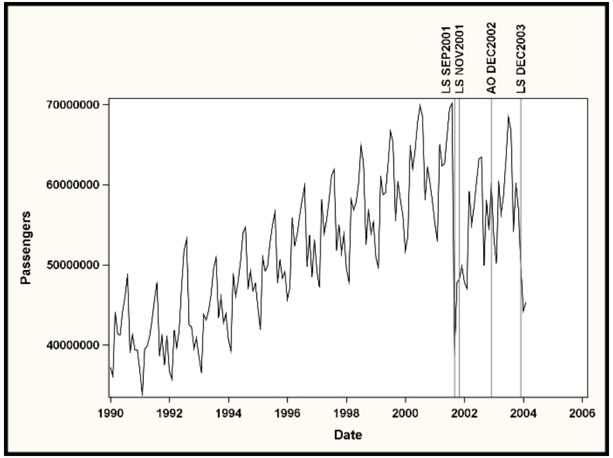


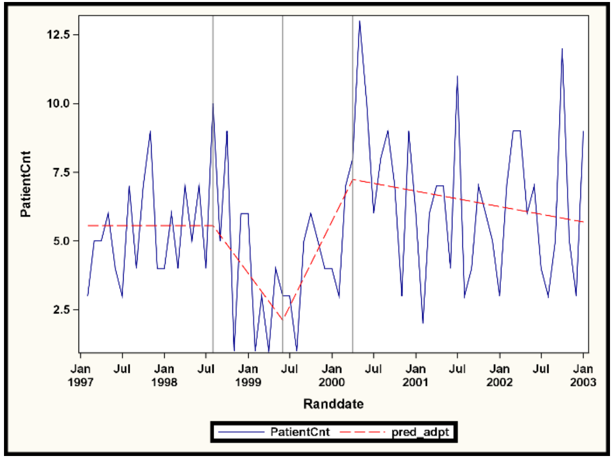
Can events and changes in the course over time be automatically detected?

```
Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models
```

Automatically Detect Breakpoints and Outliers

Use machine learning methods to identify time points in your data where the course over time deviates from „normal“ behavior.





Detecting shifts and pulse events in your data with ARIMA Models.

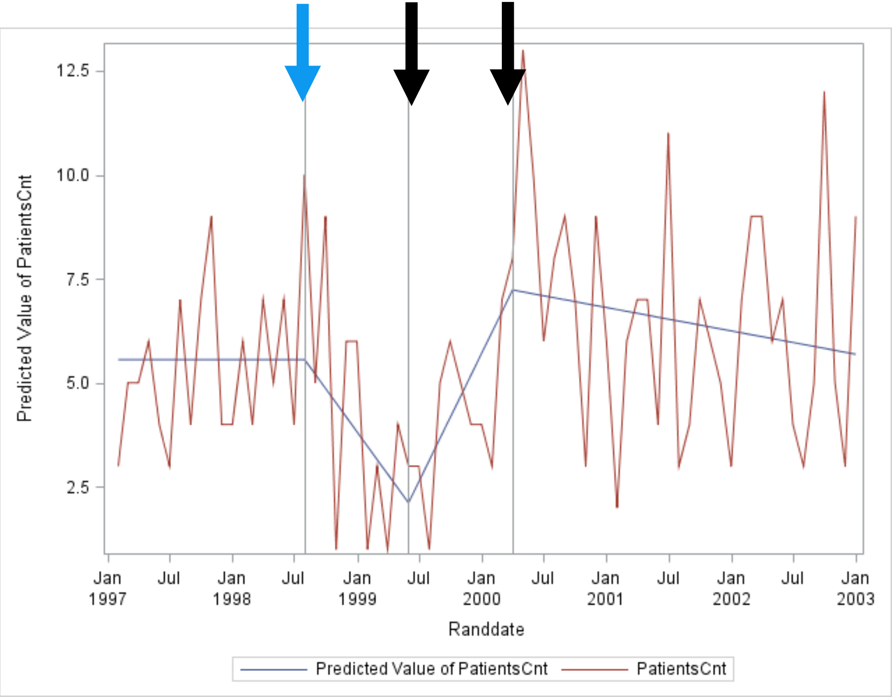
Use multivariate regression splines to identify breakpoints over time.

What happened in my clinical trial at certain points in time?

One Patient has a severe adverse event. Possible Relation to Study medication.

Study Meeting to announce „not related to treatment“

Recruitment recovers

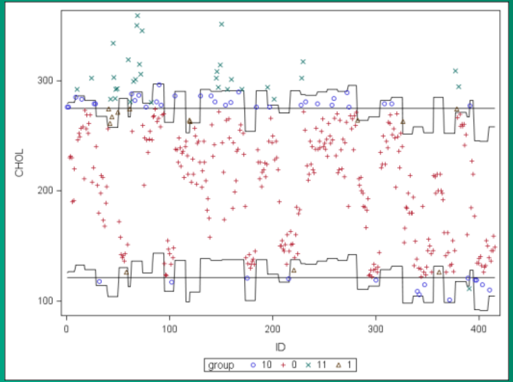


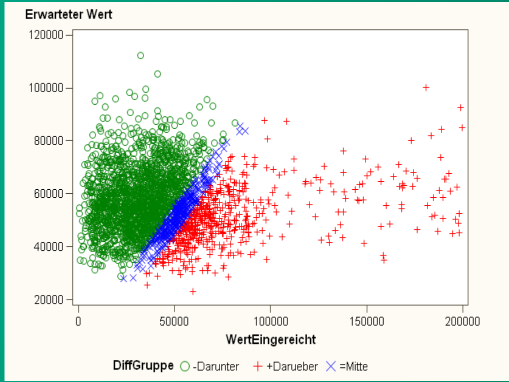
Data Science in Action: #3

Proving a reference value that considers all available co-information

Can analytics help me to reduce the "Yes, but … " sentences in my business discussions?

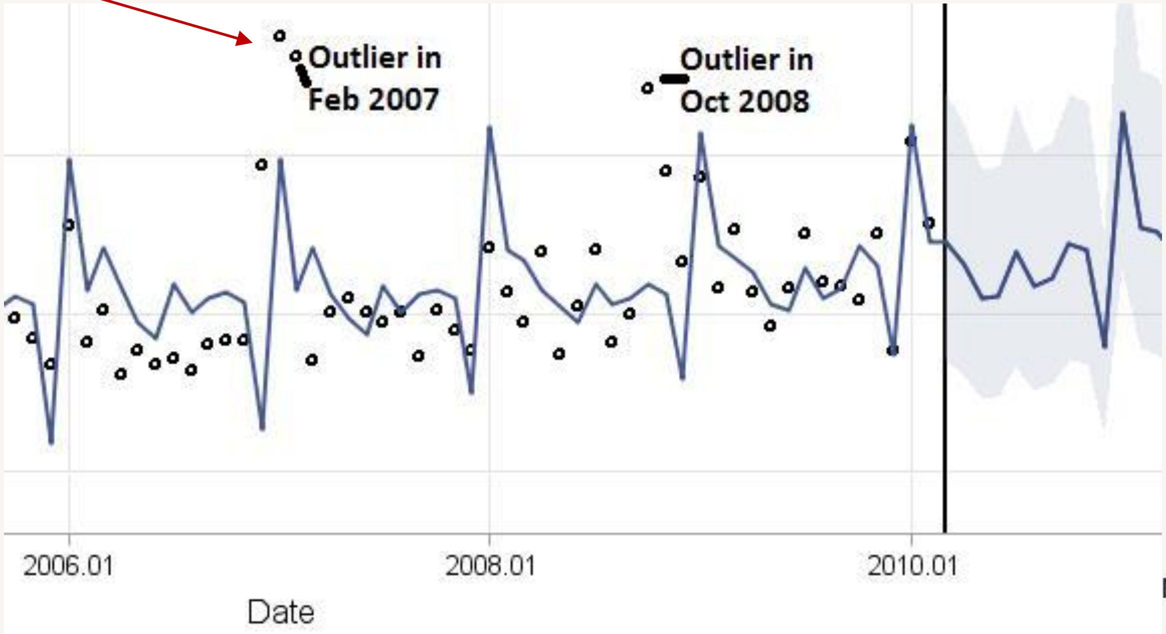
Linear Regression
Decision Trees
Time Series Analysis



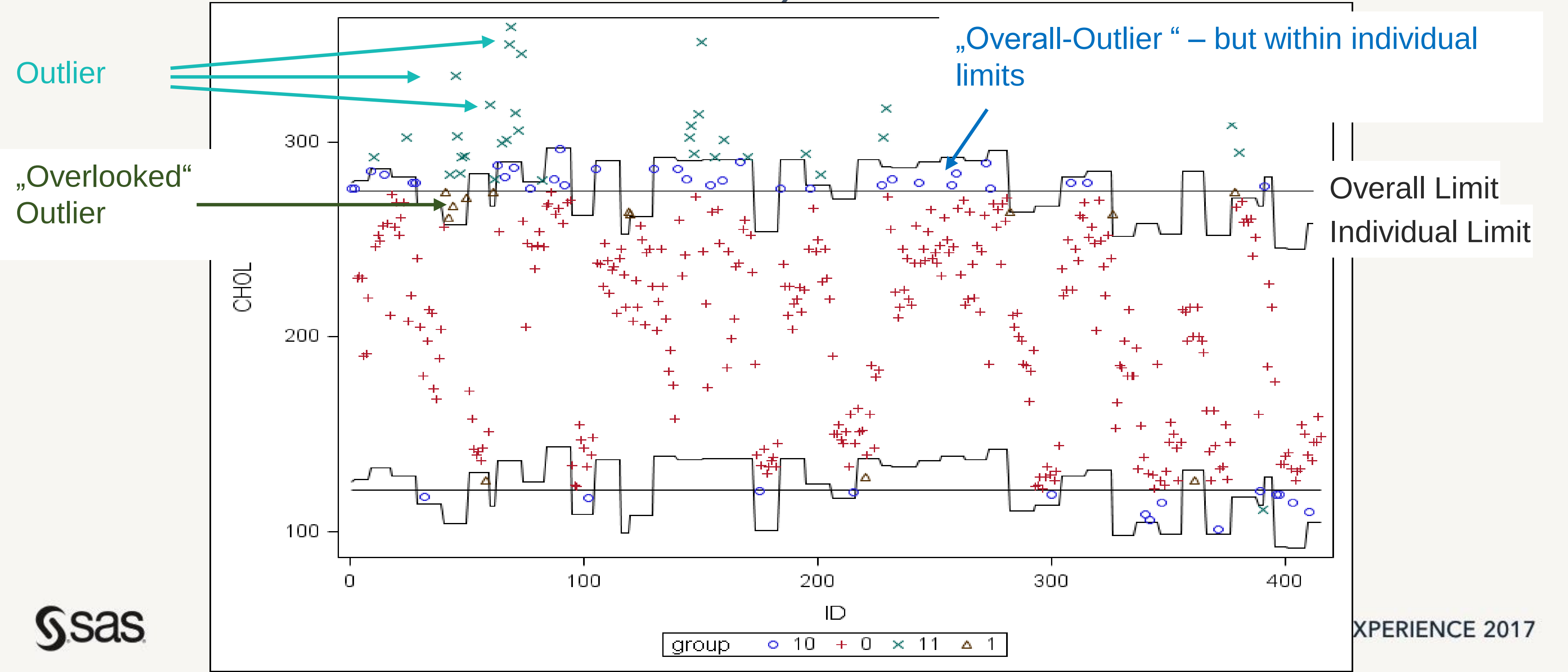


„Yes, but ....
... in January we usually have a lot of events.“

Model recognizes that this value in January is NOT an outlier



„All values larger than x are outliers! - Really?“

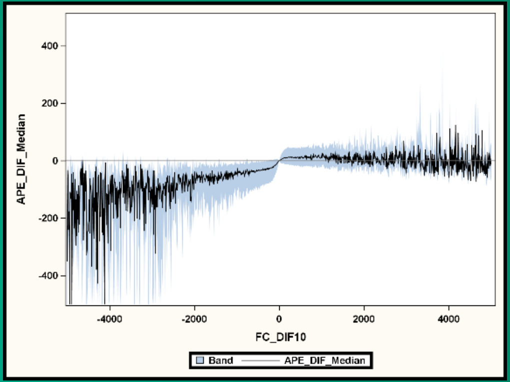


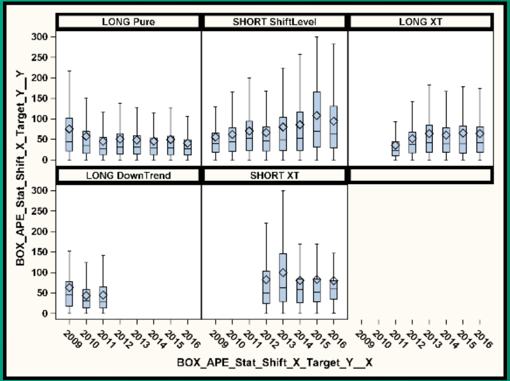
Data Science in Action: #4

Explaining Forecast Errors and Deviations

Do the demand planners really improve forecast accuracy with their manual overwrites?

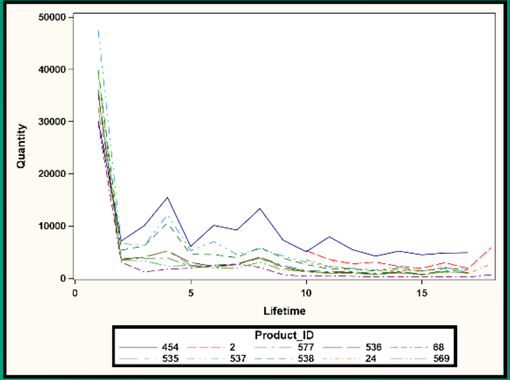
```
Linear Regression
Quantile Regression
Descriptive Statistics
```



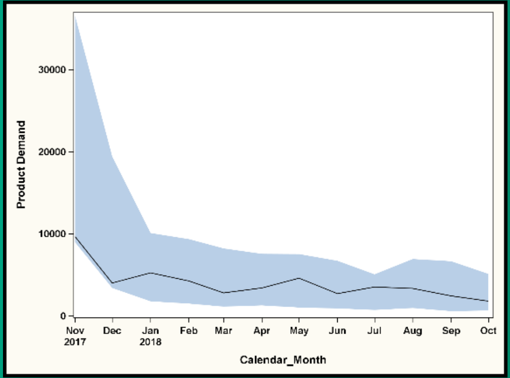


Data Science in Action: #5

Forecasting the Demand for New Products



Can the expected demand of products that are introduced only right now be estimated for forecast planning?



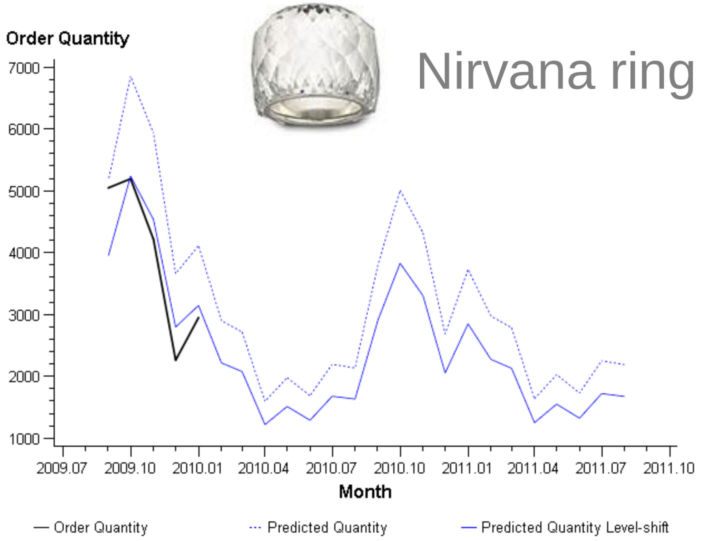

```
Poisson Regression
Cluster Analysis
Similarity Search
```

NOVELTY FORECASTING

- Training data from previous collections
- Generalized linear model

- Predictors
  - Product attributes
  - Time-dependant influence factors
  - Number of shops
  - Actual order intake
  - Actual sell-through

Nirvana ring

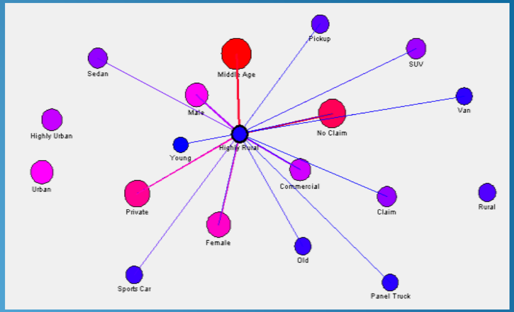


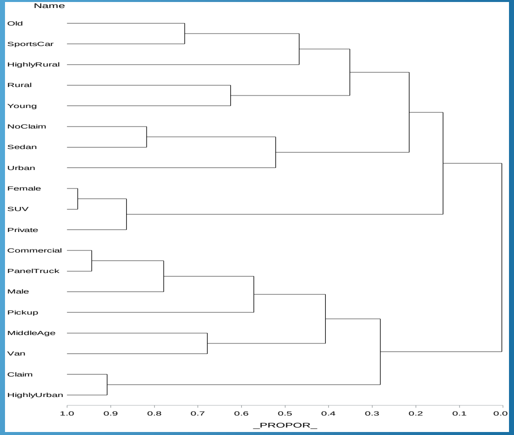
Data Science in Action: #6

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?

Unsupervised machine learning methods: association analysis variable clustering



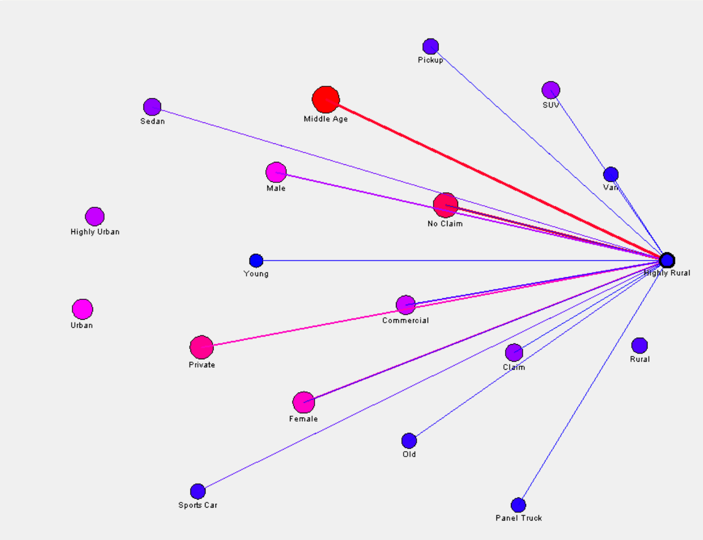


Make your Data Talk to You!

- Data from Car Insurance with 6 properties per customer

Variable	Feature
AGE	YOUNG, MIDLIFE, OLD
GENDER	MALE, FEMALE
DENSITY	HIGHLY URBAN, URBAN, HIGHLY RURAL, RURAL
CAR_TYPE	VAN, SPORTS CAR, SUV, SEDAN, PICK UP
CAR_USAGE	PRIVATE, COMMERICIAL
CLM_FLAG	CLAIM, NO CLAIM

- Use unsupervised machine learning (association analysis) to uncover relationships between different properties.



Try it! Transpose your data how you usually would not do it.

One-Row-Per-Subject

POLICYNO	CLM_FLAG	CAR_USE	CAR_TYPE	AGE	GENDER	DENSITY
160	No	Private	Sedan	60	M	Highly Urban
24836	No	Commercial	Sedan	43	M	Highly Urban
28046	No	Private	Van	48	M	Urban
28960	No	Private	SUV	35	F	Highly Urban
40933	No	Private	Sedan	51	M	Highly Urban
55277	No	Private	SUV	50	F	Urban
63212	Yes	Commercial	Sports Car	34	F	Highly Urban
69651	No	Private	SUV	54	F	Highly Urban
88070	Yes	Private	Sedan	40	M	Urban
93553	No	Commercial	SUV	44	F	Rural
127444	Yes	Commercial	Van	37	M	Highly Urban
141509	Yes	Private	SUV	34	F	Highly Urban
145326	No	Commercial	Van	50	M	Rural
146809	Yes	Private	Sports Car	53	F	Urban
148250	No	Private	Sedan	43	F	Rural
157851	No	Commercial	Van	55	M	Urban

Multple-Row-Per-Subject
Key-Value Table

POLICYNO	Feature
160	Highly Urban
160	No Claim
160	Sedan
160	Private
160	Male
160	Old
24836	Highly Urban
24836	No Claim
24836	Sedan
24836	Commercial
24836	Male
24836	Middle Age

Men Do Not Drive Sports Cars?

Rule 278 shows that sports cars are only driven in 2.54% of the cases by men, whereas this was expected in around 46% of the cases.

index	RULE	_LHAND	_RHAND	COUNT	SUPPORT	EXP_CONF	CONF	LIFT
267	Commercial ==> Sports Car	Commercial	Sports Car	200.00	1.94	11.44	5.28	0.46
268	Rural ==> Claim	Rural	Claim	102.00	0.99	26.66	6.52	0.24
269	Claim ==> Rural	Claim	Rural	102.00	0.99	15.18	3.71	0.24
270	Young ==> Highly Urban	Young	Highly Urban	10.00	0.10	34.93	8.33	0.24
271	Highly Rural ==> Claim	Highly Rural	Claim	32.00	0.31	26.66	6.30	0.24
272	Claim ==> Highly Rural	Claim	Highly Rural	32.00	0.31	4.93	1.17	0.24
273	Van ==> Female	Van	Female	117.00	1.14	53.82	12.70	0.24
274	Female ==> Van	Female	Van	117.00	1.14	8.94	2.11	0.24
275	Panel Truck ==> Female	Panel Truck	Female	40.00	0.39	53.82	4.69	0.09
276	Male ==> SUV	Male	SUV	99.00	0.96	27.98	2.08	0.07
277	SUV ==> Male	SUV	Male	99.00	0.96	46.18	3.43	0.07
278	Sports Car ==> Male	Sports Car	Male	30.00	0.29	46.18	2.54	0.06

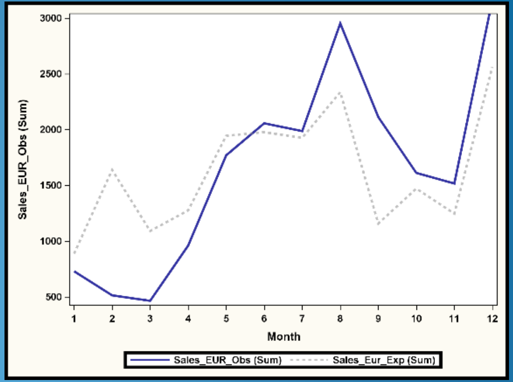
- This might indicate a situation that for the customer base, sports cars are really predominantly driven by women.
- It could be a trigger to an investigation of the quality status of your data.
- A business interpretation could be that in a family, the sports car is the 2nd or 3rd car that is registered in the wife's name for financial reasons.
- The competitor is offering a policy to men for a much more attractive price.

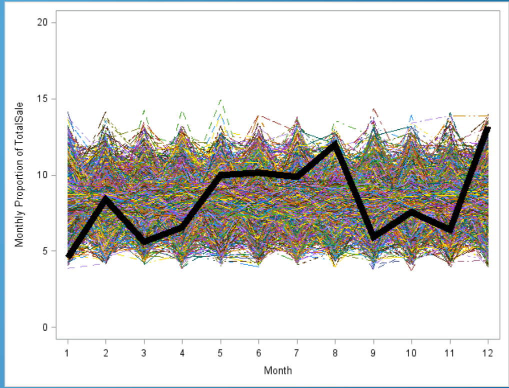
Data Science in Action: #7

Checking the Alignment with Predefined Pattern

Which customers show a behavior that is far from what you expected?

```
Chi2 independency test
Benford's law
Time Series Similarity
```

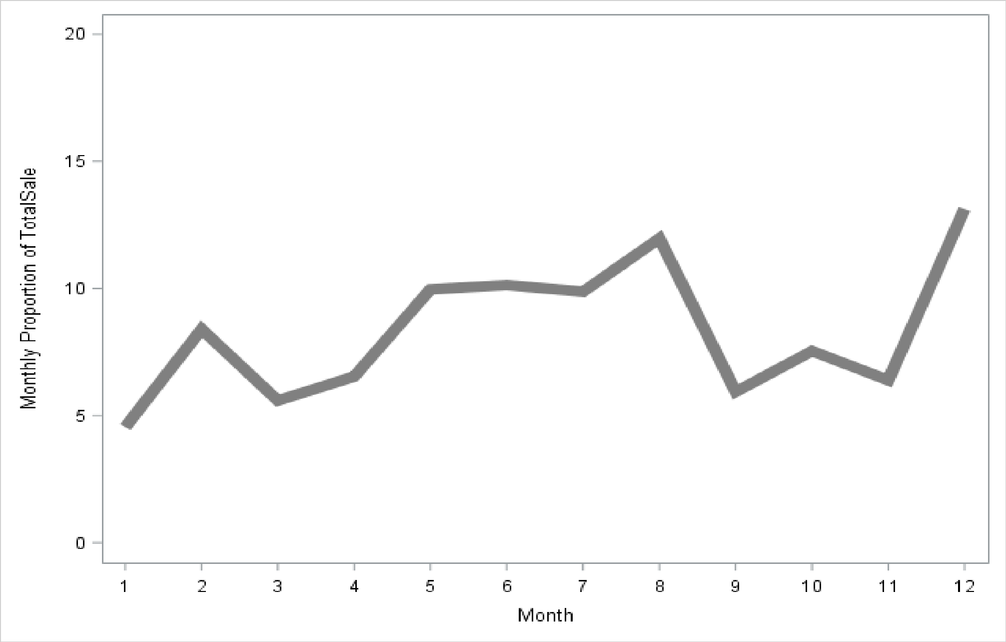




Which of my sales representatives do not follow pre-defined pattern?

The demand for sub-contractors for a company in the catering business varies over the calendar year.

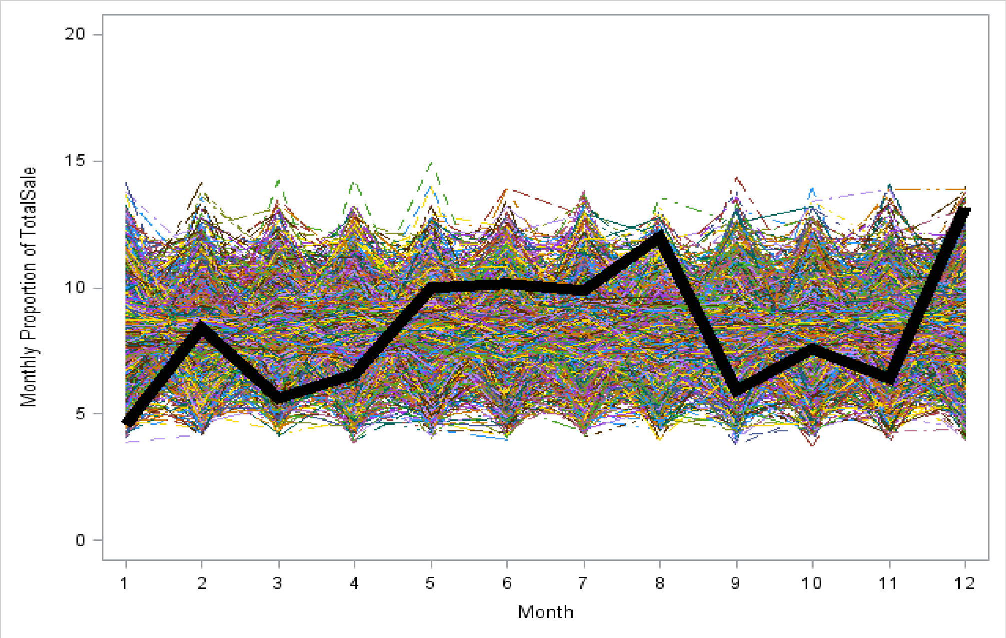
Sales Persons are forced to close such sub-contracts following the seasonal demand pattern.



Looking at the individual seasonal pattern per sales person does not help

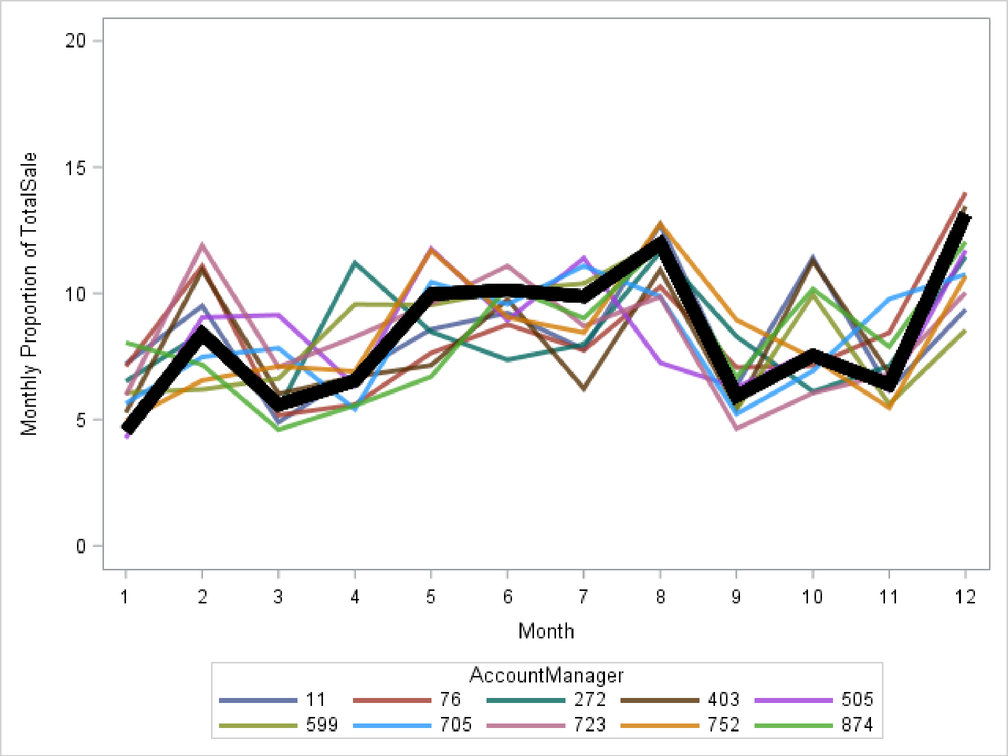
No clear picture.

Infeasible to review all individual lines manually.



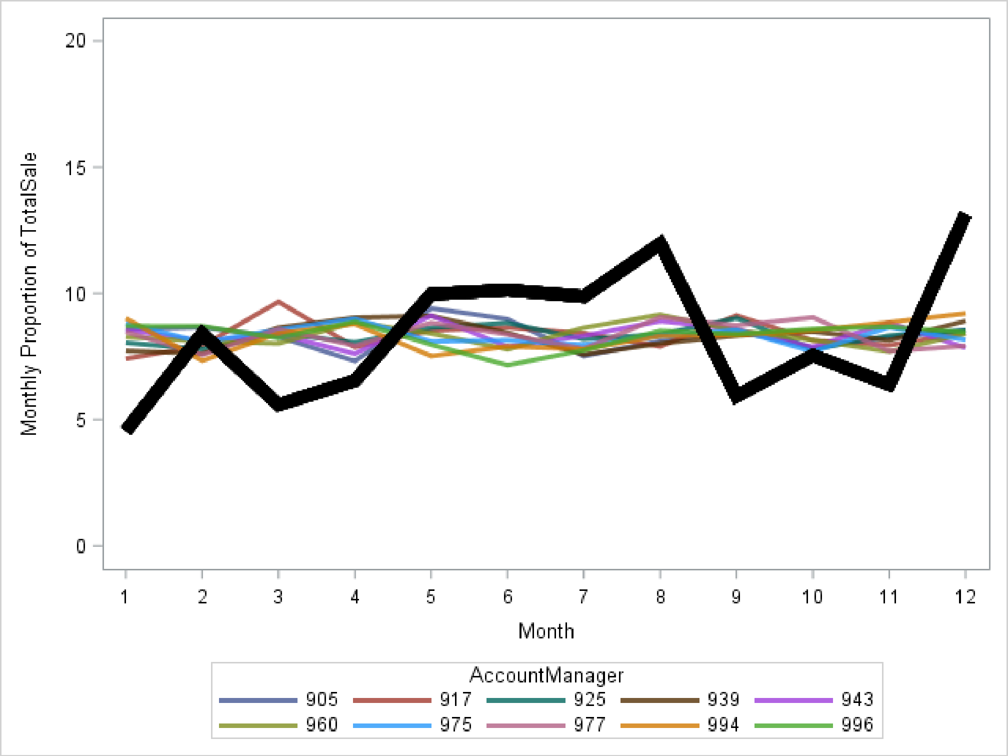
Use Analytical Methods to Rank Your Sales Persons (1)

Top 10 sales persons adhering to the pattern



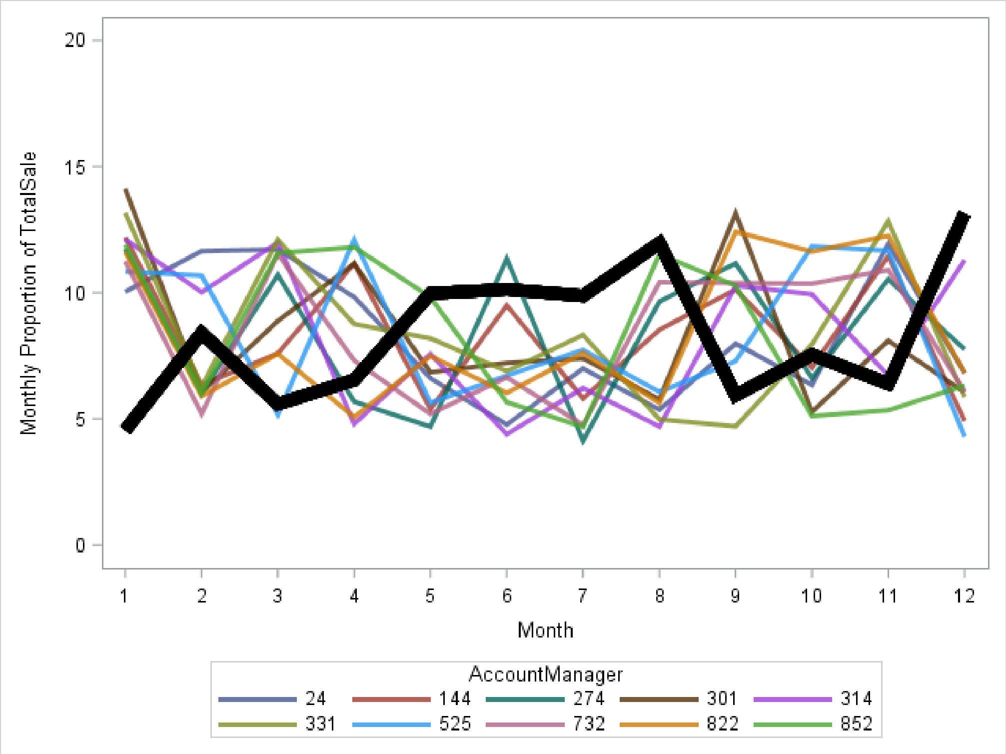
Use Analytical Methods to Rank Your Sales Persons (2)

10 sales persons without seasonal variation

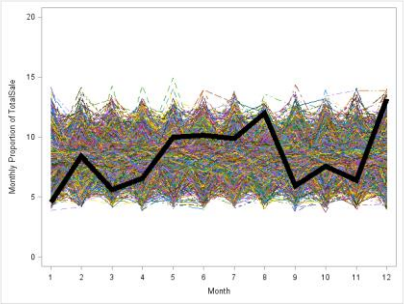


Use Analytical Methods to Rank Your Sales Persons (3)

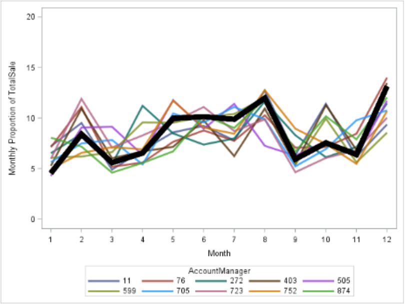
10 sales persons that work “against” the predefined pattern

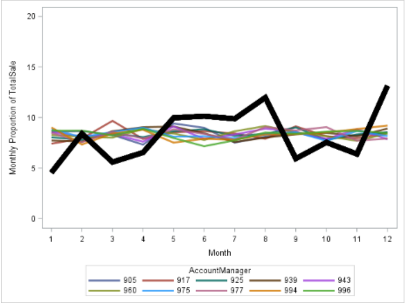


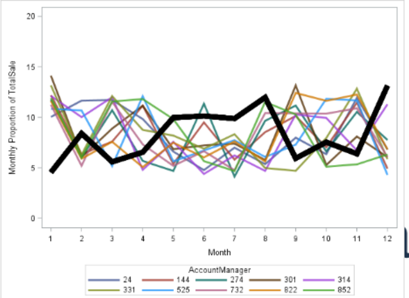
Analytics helps me, to get clearer picture!



Analytic Methods







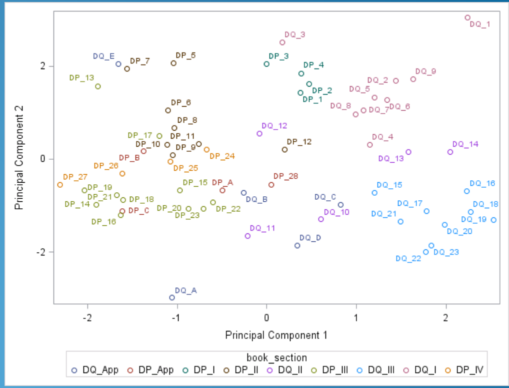
From noise

to managable segments

Data Science in Action: #8

Topic Search Documents and Clustering

Can I automatically find clusters of documents with similar content?

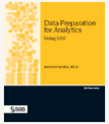


Missing Values	10	11	A																	
Erzeugen des Analytic-Marts	10	11	14	15	16	18	19	20	21	22	23	24	25	26	27	28	9	A	C	B
Data Origin und Data Management	5	13	17	B																
DQ Case Studies	1																			
Fachliche Konzepte	1	2	3	4	12	2	3	4	5	6	7	8	9	12						
DQ mit Analytik und SAS	13	14																		
Data Quality Simulationen	15	16	17	18	19	20	21	22	23	C	D									
Analytic Data Mart Structures	6	7	8	E																

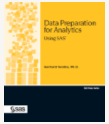
Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

Can I detect similar chapters, without having to read Gerhard's' books ☺?

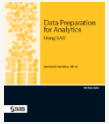
Topic > +access,+file,+text,+relational,+relational database



PAGE 104 Data Preparation for Analytics Using SAS Chapter 13: Accessing Data PAGE 103 Part 3 Data Mart Coding and Content Chapter 13 Acces
Transposing One- and Multiple-Rows-per-Subject Data Structures 115 Chapter 15 Transposing Longitudinal Data 131 Chapter 16 Transformations of
Chapter 17 Transformations of Categorical Variables 161 Chapter 18 Multiple Interval-Scaled Observations per Subject 179 Chapter 19 Multiple Catego



PAGE 38 Data Preparation for Analytics Using SAS Chapter 5: The Origin of Data PAGE 43 Part 2 Data Structures and Data Modeling Chapter 5 The
Models 45 Chapter 7 Analysis Subjects and Multiple Observations 51 Chapter 8 The One-Row-per-Subject Data Mart 61 Chapter 9 The Multiple-Rows-p
Data Structures for Longitudinal Analysis 77 Chapter 11 Considerations for Data Marts 89 Chapter 12 Considerations for Predictive Modeling 95 Introdu



PAGE 178 Data Preparation for Analytics Using SAS Chapter 17: Transformations of Categorical Variables PAGE 177 Chapter 17 Transformations
Introduction 17.2 General Considerations for Categorical Variables 162 17.3 Derived Variables 164 17.4 Combining Categories 166 17.5 Dummy Codir
Multidimensional Categorical Variables 172 17.7 Lookup Tables and External Data 176 17.1 Introduction In this chapter we will deal with transformatio

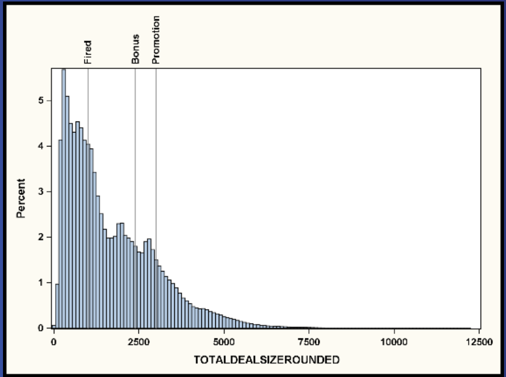
40 Data Quality for Analytics Using SAS Chapter 3: Data Availability 41 Chapter 3: Data Availability 3.1 Introduction 32 3.2 General Considerations 32 Re
data availability 32 Availability and usability 32 Effort to make data available 33 Dependence on the operational process 33 Availability and alignment in t
of Historic Data 34 Categorization and examples of historic data 34 The length of the history 35 Customer event histories 35 Operational systems and a

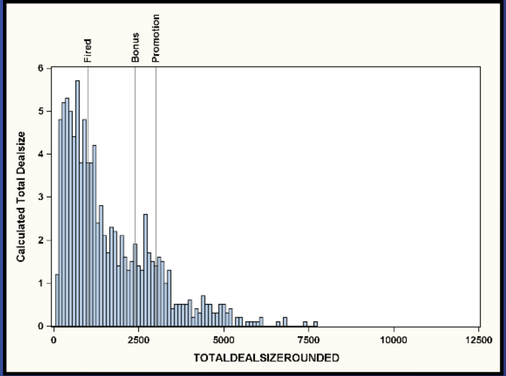
PAGE 382 Data Preparation for Analytics Using SAS Appendix B: The Power of SAS for Analytic Data Preparation PAGE 381 Appendix B The Power o
369B.1 Motivation B.2 Overview 370 B.3 Extracting Data from Source Systems 371 B.4 Changing the Data Mart Structure: Transposing 371 B.5 Data Mar
Multiple-Rows-per-Subject Data Sets 372 B.6 Selected Features of the SAS Language for Data Management 375 B.7 Benefits of the SAS Macro Langu

Data Science in Action: #9

Using Monte Carlo Simulations to Understand the Outcome Distribution

When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?



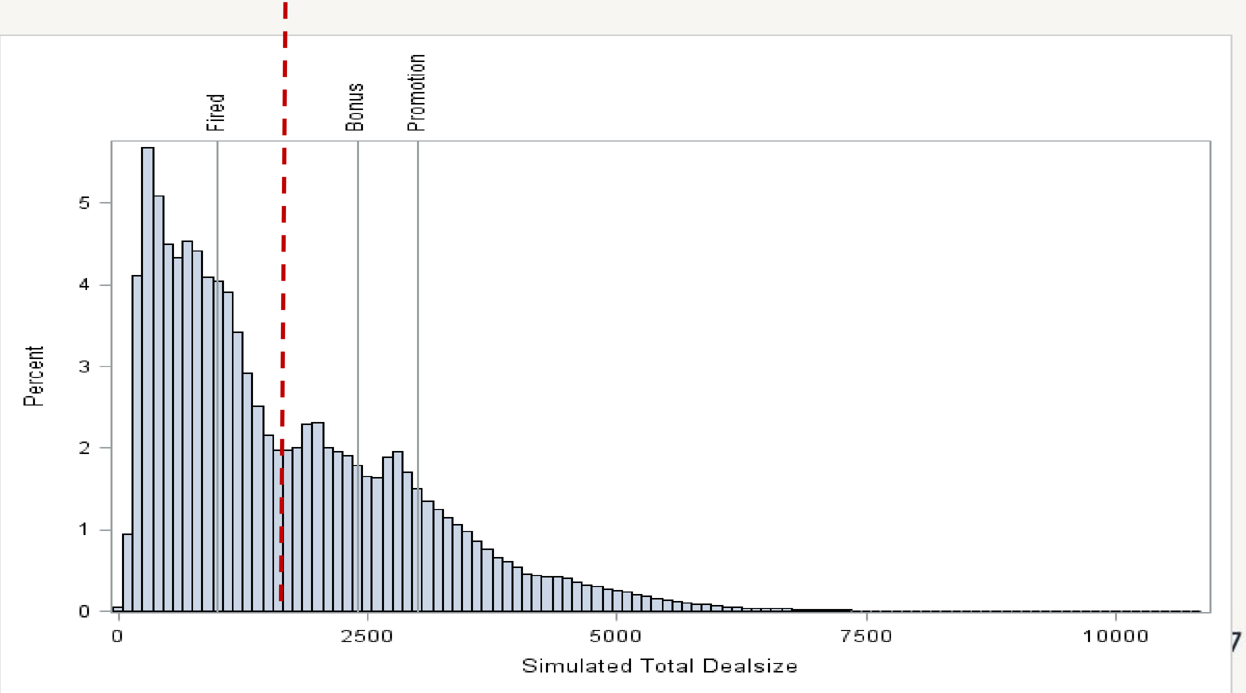


Monte Carlo simulations
Mathematical programming

Will the Sales Manager keep his Job?

ProjectID	DealSize (1000 $)	Proba- bility
1	1500	10%
2	10	65%
3	500	20%
4	50	50%
5	100	40%
6	30	90%
7	10	60%
8	150	20%
9	200	25%
10	180	10%
11	900	10%
12	750	20%
13	600	10%
14	320	20%
15	100	40%
16	50	80%
17	2000	5%
18	400	20%
19	2500	10%
20	1700	15%
21	100	80%

Weighted Average:
$ 1.661.500

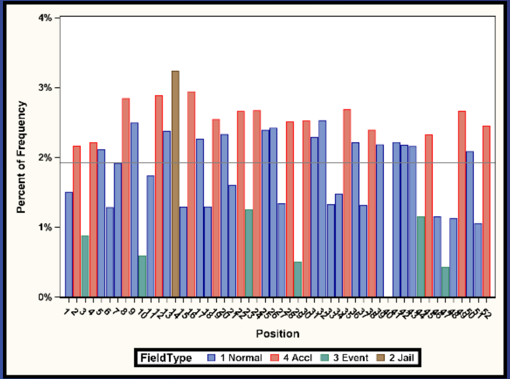


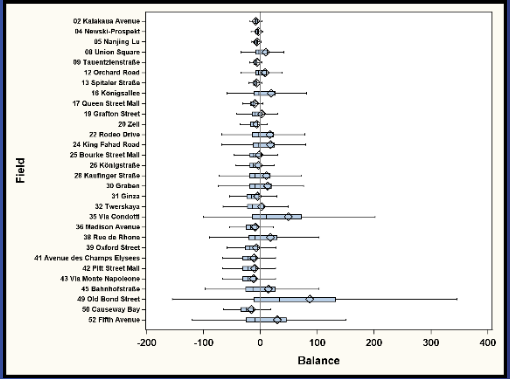
Data Science in Action: #10

Studying Complex Systems – Simulating the Monopoly Board Game

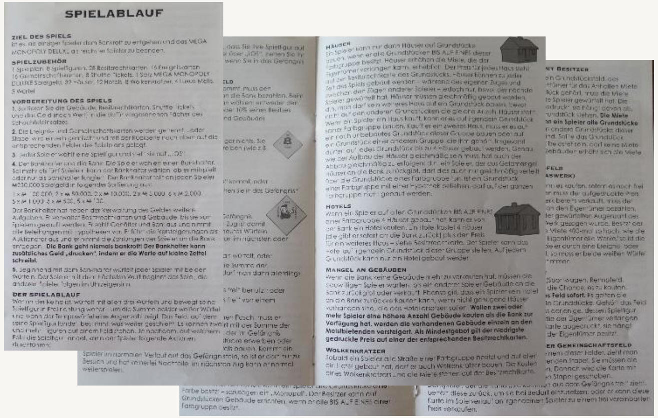
How can you simulate complex environments to get insight in the most frequent processes?

Monte Carlo Simulations

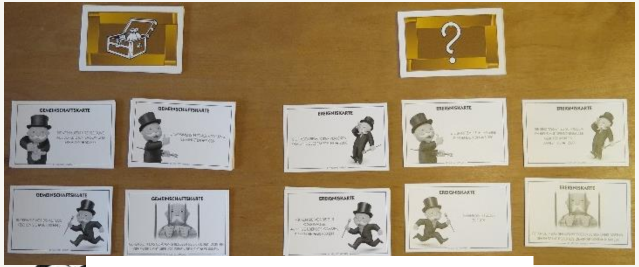




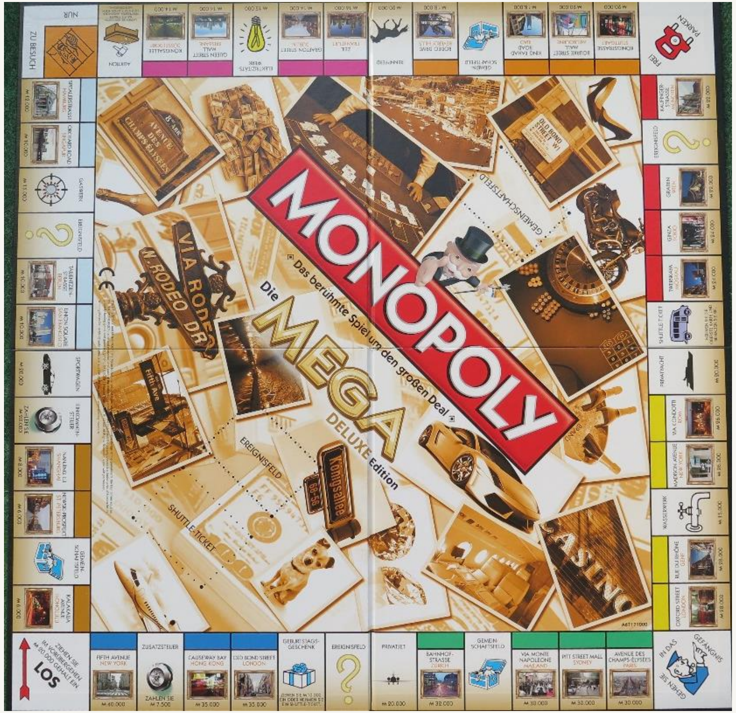
The Monopoly® board game and business life have many similarities



Set of Complex Rules



Additional Instructions



Framework of Opportunities and Events

Monetary Dimension

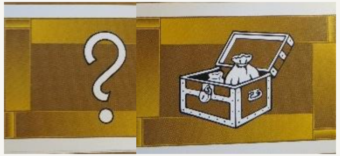
Random Components

Dynamic Component

Simulating complex processes provides insight (that you otherwise might miss)

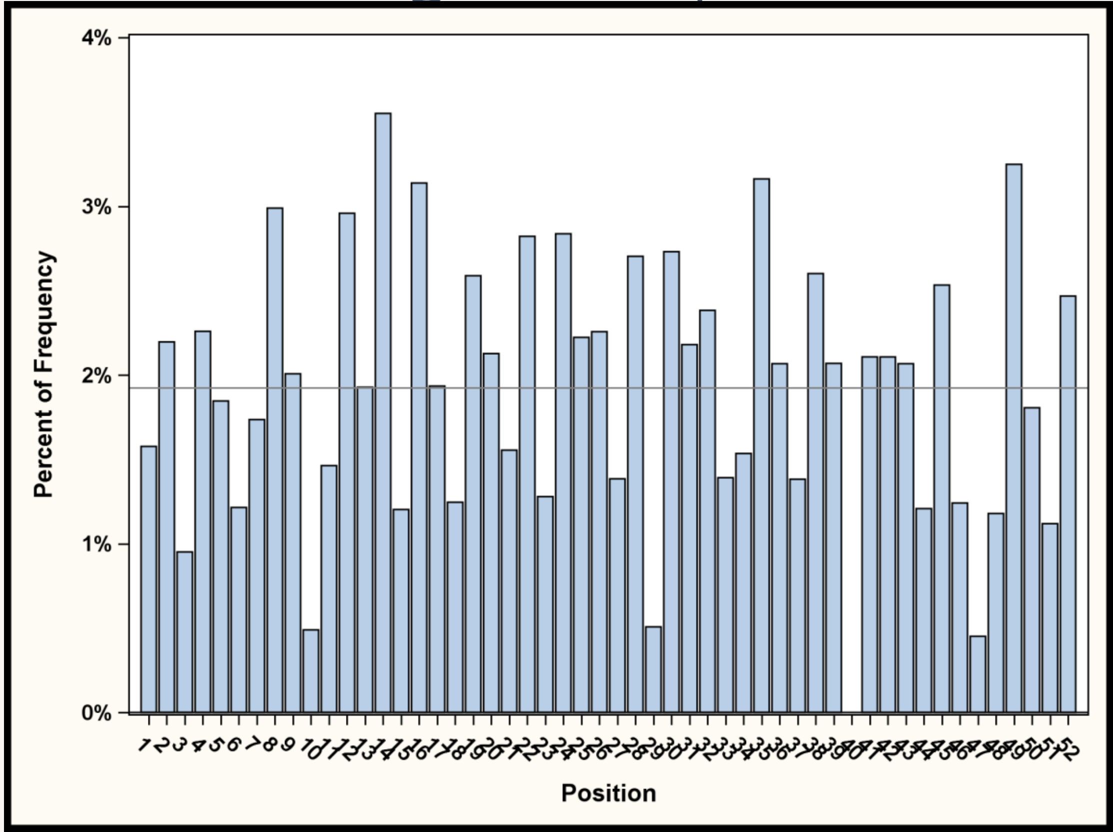
Sum of 2 Dice

Go to Jail!



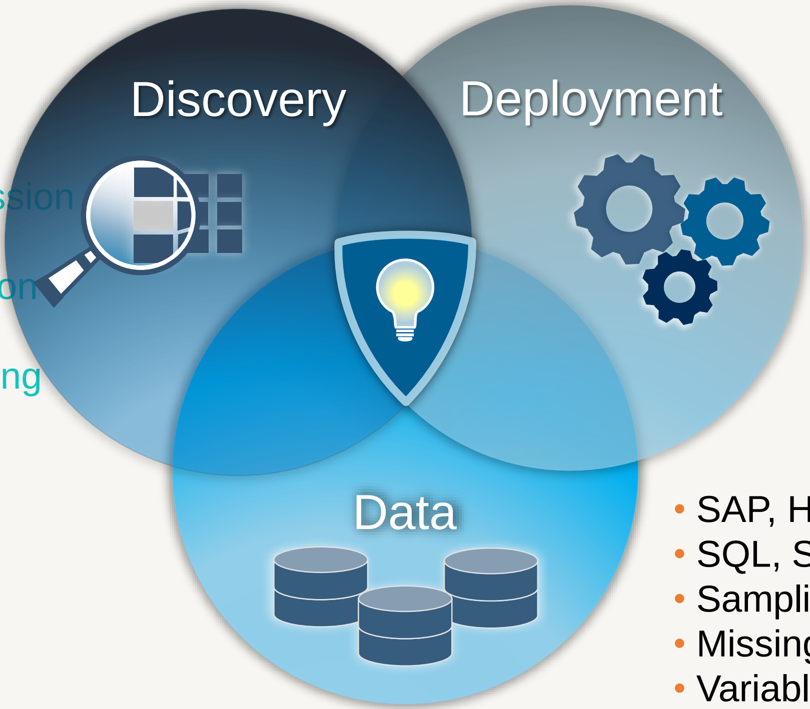
Event Fields

Accelerator Dice



Data Mining and Machine Learning with the SAS Platform

- Logistic Regression
- Linear Regression
- Generalized Linear Models
- Nonlinear Regression
- Ordinary Least Squares Regression
- Decision Trees
- Partial Least Squares Regression
- Quantile Regression
- K-means and K-modes Clustering
- Principal Component Analysis
- Random Forest
- Gradient Boosting
- Neural Networks
- Support Vector Machines
- Factorization Machines
- Network Analytics/Community Detection
- Text Mining
- Boolean Rules
- Auto-tuned Hyper-parameters



- Assess Supervised Models
- Model Management
- Deployment
- Periodic Validation
- Model-Retirement
- Retraining of Models

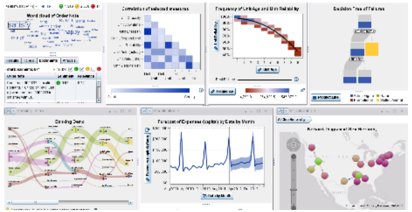
- SAP, Hadoop, Streaming, rel.DB, …
- SQL, SAS Datastep, Matrix
- Sampling and Partitioning
- Missing Value Imputation
- Variable Binning
- Variable Selection
- Transpose

Openness of the SAS Analytic Platform for different User Types

Fulfilling Individual Requirements

Office Integration

Mobile



python

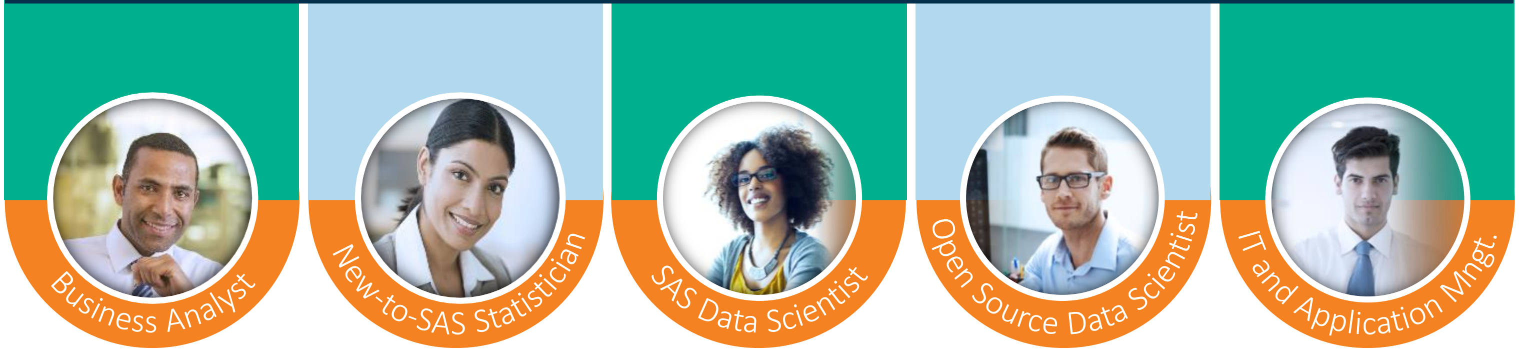
sas

REST API

Lua

Java

One Integrated Solution for Different User Types



Key Takeaways

Analytics and Data Science is there to help you!

- Get a clearer, more objective picture of your data and your analysis subjects
- Get explicit results instead of searching the needle in the haystack
- Make your data talk to you!
- Receive findings automatically instead of manually
- Do it again! – treat models as an asset and repeat your analysis

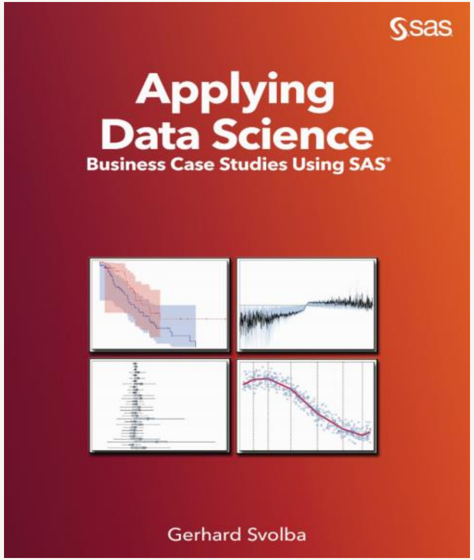
Machine Learning and Data Science are a core part of the SAS Analytic Platform

- Comprehensive set of methods – Discover and Operationalize
- Open to different user types (Coding, Point&Click, SAS, R, Python, …)

More Information

Gerhard Svolba – Principal Solutions Architect

sastools.by.gerhard@gmx.net



- Applying Data Science – Business Case Studies Using SAS, SAS Press 2017
- Eight Case Studies showing how Data Science and Analytics can be applied to provide insight into yout data and improve your business decisions
- http://www.sascommunity.org/wiki/Applying_Data_Science_-_Business_Case_Studies_Using_SAS

Further Links

- Gerhard Svolba: Mehr als linear oder logistisch – ausgewählte Möglichkeiten neuer Regressionsmethoden in SAS - Download the presentation and the paper
- Allison, P. 1995. *Survival Analysis Using SAS®: A Practical Guide, Second Edition*. Cary, NC: SAS Institute Inc.
- *SAS/STAT® 14.2 User's Guide. The LIFETEST Procedure*. http://support.sas.com/documentation/onlinedoc/stat/142/lifetest.pdf (accessed 1 March 2017).
- Kuhfeld, W., and W. Cai. 2013. "Introducing the New ADAPTIVEREG Procedure for Adaptive Regression." SAS Global Forum Proceedings. http://support.sas.com/resources/papers/proceedings13/457-2013.pdf (Paper 457-2013).